



12

24

9

TERMINOLOGY

bias
central limit theorem
cluster sampling
completion bias
convenience sampling
interviewer bias
judgement sampling
non-response bias
parameter
population
pseudo-random number
quota sampling
random number
recall/reporting bias
sample
sample proportion
sampling distribution
selection bias
self-selection bias
simple random sample
statistic
stratified random sampling
systematic sampling

INTERVAL ESTIMATES FOR PROPORTIONS

RANDOM SAMPLES AND PROPORTIONS

- 9.01 Random samples and bias
 - 9.02 Selection of samples
 - 9.03 Variability of random samples
 - 9.04 Sample proportions
 - 9.05 Parameters of sample proportions
 - 9.06 The central limit theorem
 - 9.07 Sample proportions and the standard normal distribution
- Chapter summary
- Chapter review



Prior learning

RANDOM SAMPLING

- understand the concept of a random sample (ACMMM171)
- discuss sources of bias in samples and procedures to ensure randomness (ACMMM172)
- use graphical displays of simulated data to investigate the variability of random samples from various types of distributions, including uniform, normal and Bernoulli (ACMMM173)

SAMPLE PROPORTIONS

- understand the concept of the sample proportion \hat{p} as a random variable whose value varies between samples, and the formulas for the mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$ of the sample proportion \hat{p} (ACMMM174)
- examine the approximate normality of the distribution of \hat{p} for large samples (ACMMM175)
- simulate repeated random sampling, for a variety of values of p and a range of sample sizes, to illustrate the distribution of \hat{p} and the approximate standard normality of $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$ where the closeness of the approximation depends on both n and p (ACMMM176) 

9.01 RANDOM SAMPLES AND BIAS

Many variables involve the collection of data from very large groups. It may not be practical to use the whole group. It may be more practical to collect information from only part of the group.

IMPORTANT

For any variable or group of variables, the **population** (or **population of interest**) is the whole group from which data could be collected. It is the universal set for the data.

A **sample** is a part of the population.

In a **census**, data is collected from the whole population.

A **parameter** is a characteristic value of a particular population, such as the mean.

A **statistic** is an estimate of a parameter obtained using a sample.

A **survey** obtains the same information from each member of the sample or population. For a survey of people you would ask each person the same questions.

Example 1

A sample of twenty people waiting in an ATM queue at 7:30 a.m. were asked how much they intended to withdraw. The smallest amount was \$20, the average amount was \$78, and the greatest was \$500. Identify the population, some parameters and statistics.



Getty Images/Bloomberg

Solution

The population is the whole group that could be asked about the amount they withdraw from an ATM.

Parameters are clearly defined values from the whole population. You don't need to know the value to define it clearly.

Statistics are the values you get from the sample. The number of people (20) is not a statistic because it is not an estimate of a parameter.

The population is all the people who use ATMs.

There are 3 parameters: the minimum withdrawal, the average amount withdrawn and the maximum withdrawal.

There are 3 statistics: the minimum withdrawal (\$20), the average withdrawal (\$78) and the maximum withdrawal (\$500).

The sample size is not a parameter because it is not a population property. The size of the population is a parameter.

When you use a sample to find a statistic, you want the statistic to be as close as possible to the population parameter. You need to choose the sample so that it is representative of the population.

Using a very small sample will not usually give you an accurate representation of the population. If you used the extreme case of a sample of size 1, it is obvious that this will not give good results.

You cannot guarantee that a sample will be representative. Suppose that you wanted to find the average income of people in a particular area. Unless you checked everyone, you might miss the one person who was a millionaire. This would obviously have a big effect on the statistic from your sample.

IMPORTANT

A **fair sample** is one that is representative of the population.

A **biased sample** is not representative: it favours some section of the population.

A **random sample** ensures that every member of the population has an equal chance of being chosen.

The statistics from a fair sample are likely to be close to the parameters of the population. Those from a biased sample are unlikely to be close to those of the population.

A random sample is more likely to be fair than one chosen by other means, so statisticians prefer random samples. Unfortunately, random samples of large groups are generally difficult and expensive to obtain.

There are many sources of bias in statistical studies. Investigations involving opinion or feelings are more likely to involve bias than those where you are making measurements.

- **Selection bias** arises from the choice of the sample. This is best avoided by using a random sample.
- **Design flaw bias** arises from faults in the design of the study. Use objective measures wherever possible. If opinion or other subjective measures are required, focus clearly on the question. For example, rather than asking ‘Do you support the Liberal party?’, ask ‘If an election were held this Saturday, which party would you be most likely to vote for?’

Bias can arise during data collection from differences in the way that data is collected.

- **Interviewer bias** can occur through differences in the way that different interviewers seek information. This can be minimised using standard questions and question options.

In medical trials, special procedures are used to reduce bias. For example, in a double-blind trial some patients are given a drug and others are given a placebo, but the doctors and patients are not told which. This reduces bias caused by the doctor or patient knowledge of treatment.

- **Recall/reporting bias** arises when knowledge of the outcome of one answer affects recall or reporting of the answer. For example, a question about voting intention could affect reporting of past voting practice. Even asking about something that happened the previous week could give false results due to inaccurate recollection.
- **Completion bias** occurs when surveys are incomplete. This can mean that later questions are biased because the questionnaire or interview is abandoned. Surveys should be as short as practical. Longitudinal studies use similar surveys of the same group over an extended period. Completion bias is a big problem, as lost subjects could have a systematic effect on outcomes. A longitudinal study of rural employment in remote areas would be affected by the loss of people who moved away.

Some people will refuse to answer sensitive questions. This is minimised by avoiding questions of potential embarrassment. For example, rather than asking ‘how much do you earn in a week?’ put it as ‘tick the box that shows your income category’.

- **Non-response bias** occurs when some subjects do not take part in the survey. You should be sure that this does not systematically affect results. The most extreme example of non-response bias is a **self-selected sample**. An example of a such a sample is the group of people who respond to a media poll, such as a TV program that asks people to respond ‘Yes’ or ‘No’ to a question by ringing up, texting or logging on to a site.

Even if a survey is framed and conducted with minimum bias, bias can still be introduced by inappropriate analysis or reporting of results.

○ Example 2

In each of the following cases, state whether or not the sampling method is fair, and if it is biased, state the kind(s) of bias.

- a An interviewer outside a supermarket on Saturday morning asked people going in: ‘Do you prefer *Razzle* dishwasher detergent or an inferior brand?’
- b 2000 mobile phone numbers were telephoned at random and people answering were asked: ‘What kind of dishwasher detergent do you use?’
- c 200 people are chosen at random from the electoral roll of a ‘litmus-test’ electorate and interviewed at home on a Sunday morning. They are asked ‘If an election were held tomorrow, who would you vote for?’ Anyone who wasn’t home was contacted later at a follow-up that evening. Altogether, 190 people were successfully contacted and only 10 refused to answer. They were put into the ‘don’t know’ category.

Solution

- a Those interviewed were available at a particular place and time only. They were asked a leading question.
- b Mobile phones are still not common among elderly people. The question seems fair, but some people would just hang up.
- c Within the electorate, the method and follow-up gives as close to a random selection as practical. However, just because this electorate has gone with the government in the past doesn’t mean it will in the future.

The survey is biased, with both selection bias and design flaw bias.

While the design of the question is good, there is some selection bias and there is likely to be some non-response bias.

The method is fair within the electorate, but if the intention is to predict the result in terms of government, there is selection bias.

Example 2 shows that it is virtually impossible to avoid some bias in a survey, and that this may even be true for a census because of non-response bias. The Australian Bureau of Statistics (ABS) has the legal power to demand answers for its surveys, but this does not guarantee that people will give genuine answers. For important surveys, they use advanced statistical methods to ensure that the final analysis is as free of bias as possible.

INVESTIGATION

Political polls

Polls of voting intentions generally have samples of 1000–3000 voters from a large part of the state or nation. Examine some recent polls to find how they have selected the people they use, and how they ensure as much randomness as possible.

Such polls often have the collected data analysed to produce a ‘two-party preferred’ result. In this kind of analysis, what do they do about people who did not answer or who gave an answer of ‘don’t know’? Is the method reasonable?

What do they do about people who said they would vote for one of the candidates or parties not selected as one of the two parties? Is this a reasonable method?

Historically, one of the most famous political polls was the ‘Readers Digest’ poll of 1936. It predicted a 60–40 massive loss for the US presidential candidate who actually won a short time later in a 60–40 landslide. This poll was sent out to 10 million voters. What went wrong with this survey?



Shutterstock.com/PRANAV/ICRA

It is not always easy to collect data for statistics to estimate a population parameter. Sometimes it is better to use a sample to find a *related* statistic that can be used in combination with other statistics to obtain the desired information.

○ Example 3

Which statistic is easier (and cheaper) to collect?

- A The proportion of full-time workers who are women.
- B The proportion of women who are full-time workers.

Solution

To find an unbiased sample of full-time workers might prove difficult, as people work in so many different industries and have such different work hours. However, it is relatively easy (and cheap) to obtain an unbiased sample of women by using the electoral rolls.

Write the answer.

B is the easier (and thus cheaper) statistic to collect.

EXERCISE 9.01 Random samples and bias

Concepts and techniques

- 1 **Example 1** Identify the population, some parameters and some statistics for each of the following.
- a The weights of 5 meat pies produced at a pie factory were 105 g, 110 g, 98 g, 101 g and 102 g. The quality control officer also found that all pies were of even colour.
 - b A feedback sheet left in guest rooms at a hotel had 2 questions about room service (each on a 5-point scale of Very good to Very poor). The responses of 10 guests were as follows:
Food quality: Very good 3, Good 4, Average 1, Poor 1, Very poor 1
Service speed: Very good 3, Good 3, Average 1, Poor 2, Very poor 1
The manager told kitchen staff that they had scored a rating of 'only 3.7' but this was 'better than the service staff'.
 - c The numbers of passengers on 10 successive 55-seat buses at Holland Park bus station were 15, 30, 45, 20, 20, 25, 46, 48, 16 and 32.
 - d Thirty people arriving at Cairns airport to catch flights were asked how long it had taken them to reach the airport. Five said less than 10 minutes, 17 said between 10 and 20 minutes, 6 said between 20 and 30 minutes and the other 2 said between 30 and 40 minutes.
 - e A large supermarket kept records of checkout operators. From 15 shift records selected at random, there were 4 operators with more than 5 errors, 3 operators with incorrect till totals and 2 operators who had processed less than \$10 000 in their shift.



Alamy/Bleed Images

- 2 **Example 3** Which statistic is easier to collect?
- A The proportion of male drivers who have serious car accidents.
 - B The proportion of drivers in serious car accidents who are male.
- 3 Which statistic is easier to collect?
- A The proportion of timber workers who make Work Cover claims.
 - B The proportion of Work Cover claims made by timber workers.
- 4 Which statistic is easier to collect?
- A The proportion of serious injuries to children incurred on school playground equipment.
 - B The proportion of children using school playground equipment who are seriously injured on the equipment.

Reasoning and communication

- 5 **Example 2** In each of the following cases, state whether or not the sampling method is fair, and if it isn't, state the kind(s) of bias.
- a A hardware store wants to know if it would be worthwhile staying open for longer hours.
A survey placed on the counter asks customers to tick the times they are likely to shop in the store from a list.
 - b A student in Year 12 goes to Year 8 classes to investigate the amount of pocket money that they receive from their parents. He asks students in the classes to tell him how much pocket money they got last week and writes down the responses from each.
 - c A reality TV show eliminates one contestant each week by having people SMS their choice of who gets eliminated to a particular number each week. They have the system set up so that only one vote is accepted from each mobile number.
 - d A council needs to establish a new landfill site for rubbish as the old one is almost full. Some councillors urge the council to use 'people power' to decide the new site by asking residents close to each of the possible sites for their opinion of the best site. It is proposed that people attending public meetings near each site will be asked to vote on the suitability of the nearby site.
 - e A food critic goes to different restaurants and tries a selection of dishes on the menu before rating the restaurants with 1–10 scales on ambience, presentation, quality of cooking and variety of the menu. The critic is very well-known in the area and his opinion is valued highly. Reviews of three restaurants are published the following week.
- 6 A school surveys its students at the beginning of the year to determine the amount of part-time work they do.
- a What is the population for this survey?
 - b What are the parameters?
 - c What problems may this survey have in determining the parameters?
- 7 A high school hall is set up for Year 12 examinations, with students seated alphabetically in rows of 25 from front to back with 11 desks in each row. Some desks are vacant due to student illness. Students are selected to provide feedback on the conduct of the exam. Every 15th student is chosen, starting with the 3rd desk in the front row and working across to the right, then back to the left in the next row, and so on until 10 students are selected. Comment on the method of sampling.
- 8 Here are some ways a student proposes to collect a sample of students in a school for interview. State the bias that may be present for each method, and select the one that you think is the fairest.
- a Ask everyone in your class.
 - b Ask the first 80 students who walk into the resource centre.
 - c Ask all students in your year level.
 - d In school assembly, announce that there will be a poll and ask the first 70 students who volunteer to do the survey.
 - e Obtain an alphabetical list of all the students at the school and ask every 20th student on the list.
 - f Leave a 'nomination sheet' in the resource centre and ask only those people who write their names on it.
 - g Ask for 5 volunteers from every form class in the school.
 - h Ask 1 in every 15 students from each year level in the school.
 - i Call a meeting of all interested students and ask all the people who attend the meeting.
 - j Wait at the entrance of the school and ask the first 100 students who arrive after 7:30 a.m.

9.02 SELECTION OF SAMPLES

However well a study or survey is designed, administered, analysed and reported, if a poor sample is chosen, the results are likely to be poor. The statistics will probably be bad estimates of the parameters.

The simplest way to obtain a random sample is to number the population in some way and then use **random numbers** to select the sample. You can choose to use a numbering system that is already present in the population. For example, you could use serial numbers of manufactured goods or identification numbers of the target population, like student numbers for school students. For very large populations, such as ‘all Australians’, it could be difficult to assign numbers.

Even if you do have a numbering system, you then need random numbers. Tables of random numbers produced from random physical phenomena are available and are commonly used when truly random numbers are required. Some hardware random number generators are also available.

The short table of 2-digit random numbers on page 468 shows what you might expect, although books of random numbers generally have at least 4 digits. If you need fewer digits, it is usual to use the first ones in the number. If you want more digits, two or more numbers are taken together.

Example 4

- Use the two-digit random number table on page 468 to randomly select 6 numbers from 87 to 524.
- CAS** Select 6 random numbers from 87 to 524.

Solution

- Start at, say, the 16th column in the 11th row.

The first two numbers are 24 70.

04 92 03 87 51	08 13 11 48 36	98 73 32 94 11	01 78 95 19 70	13 84 91 57 67
05 04 13 40 88	75 68 99 63 19	56 69 99 33 68	24 70 05 25 64	42 41 85 04 88
30 64 49 26 22	93 66 84 39 90	57 91 05 63 53	86 05 39 32 61	67 10 68 26 73
16 02 93 88 42	32 97 19 48 39	27 00 17 29 98	95 33 02 15 35	84 54 88 77 88
90 72 79 41 71	30 19 99 89 25	18 77 55 49 03	75 26 66 89 31	45 75 85 95 16
99 31 34 95 97	50 56 14 09 36	63 23 12 58 28	64 16 96 92 62	73 96 99 48 21
55 38 06 44 27	29 38 61 58 15	66 43 42 97 45	51 03 81 16 99	06 55 69 43 88

Use the first 3 digits.

The number is 247.

Get the next number.

Next is 05 25, but 052 is less than 87, so discard it.

Get the next number.

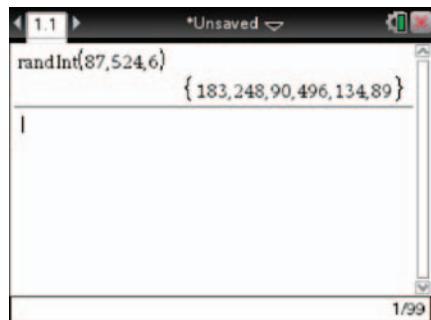
Next is 64 42, so 644.

Keep going in the same way.

6 random numbers are 247, 644, 418, 306, 492 and 229.

TI-Nspire CAS

Press **menu** 5: Probability, 4: Random and 2: Integer and put in $\text{randInt}(87, 524, 6)$. A different set of numbers is chosen each time.

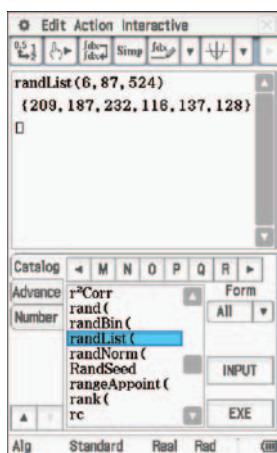


- b Write the answer.

ClassPad

Use the Main menu. Press **Keyboard**. Tap **▼** to get to the Catalog of functions. Select **randList(** and fill in the order as follows: the number of random numbers required, the lowest integer in the range, the highest integer in the range. A different set of numbers is chosen each time. You may find it useful to use the screen rotation if there are more random numbers, or simply choose fewer at the start.

6 random numbers are 183, 248, 90, 496, 134, 89.



- b Write the answer.

6 random numbers are 209, 187, 232, 116, 137, 128.

Once you have started using a table of random numbers, you usually continue any subsequent use from the place you last finished. This avoids the possibility of repeating the same set.

Using truly random numbers is very time-consuming, so it is more common to use **pseudo-random** numbers generated by a rule that produces numbers that are difficult to distinguish from random numbers. This is the method used by most modern calculators and computers, such as your CAS calculator.

INVESTIGATION

Pseudo-random numbers

A very simple pseudo-random number rule is

$$x_n = 20x_{n-1} \text{ mod } (37), \text{ where } a \text{ mod } (b) \text{ is the remainder when } a \text{ is divided by } b.$$

Start with 5 (the seed) and generate random numbers with this rule. What do you get?

How long does it take for the sequence to repeat itself?

You can use your CAS calculator by entering 5 and using the rule on the answer. Note that in the TI-Nspire, the mod function is called remain and 'ans' is replaced by the number. Pressing enter will repeat the calculation.

TI-Nspire CAS

The screenshot shows the TI-Nspire CAS calculator interface. The input field contains the command `remain(20·5,37)`. The output field shows the result 26. Below it, the next four results of the sequence are shown: 2, 3, and 23. The bottom status bar indicates 5/99.

ClassPad

The screenshot shows the ClassPad calculator interface. The input field contains the command `mod(20×ans, 37)`. The output field shows the result 5. Below it, the next four results of the sequence are shown: 26, 2, 3, and 23. The bottom status bar indicates Alg, Decimal, Real, Rad, and a small icon.

How long does it take before this sequence of pseudo-random numbers repeats?

Try the rule $x_n = 12x_{n-1} \text{ mod } 19$, starting from 7. How long does this take to repeat?

Random numbers are used extensively in encryption of computer messages, particularly payments over the internet. Obviously, the pseudo-random number used must be non-repeating over a very long sequence.

The Blum Blum Shub generator, invented in 1986, is given by $x_n = (x_{n-1})^2 \text{ mod } (pq)$, where p and q are large prime numbers. Try this rule for the prime numbers 6047 and 4723, starting from $n = 3$.

Investigate how random numbers are used for internet encryption.

There are a number of variations of random sampling that are used to avoid numbering the whole population. You can choose representatives of different groups, you can choose different groups or you can choose individuals from a group.

IMPORTANT

Simple random sampling is choosing your sample at random from the whole population.

Stratified random sampling is choosing representatives at random from identifiable groups of the population in proportion to the size of each group.

Cluster sampling is choosing **group(s)** at random from the population as the sample.

Systematic sampling is choosing representatives from the population by taking every n th to make the desired sample size.

Example 5

A distributor has 34 office staff, 23 store employees and 43 delivery drivers. How many of each should be selected to make a stratified random sample of 10?

Solution

Find the proportion of each.

The proportions are $\frac{34}{100} : \frac{23}{100} : \frac{43}{100}$, office : store : delivery.

Find the number of office staff.

$$\text{Number of office staff} = \frac{34}{100} \times 10 = 3.4 \approx 3$$

Find the number of store staff.

$$\text{Number of store staff} = \frac{23}{100} \times 10 = 2.3 \approx 2$$

Find the number of delivery staff.

$$\text{Number of delivery staff} = \frac{43}{100} \times 10 = 4.3 \approx 4$$

Write the answer.

For the balanced sample closest to 10, there should be 2 store staff, 4 drivers and 3 from the office, selected at random from each group.

If you had to have a sample of exactly 10 in Example 5, you would choose an extra person from the office. It is closest to an extra half more, so would upset the balance by the smallest amount. If you were one over and needed the exact number, you would choose the one closest to needing a half less.

When you use a systematic sample, you should start at a random place in your list. If you go past the end, you just go back to the start and keep counting.

○ Example 6

The library ID numbers of Year 8 students in a school run from 247 to 395 inclusive, since Year 8 library cards were issued after the Year 7 cards. Use the ID numbers for a systematic sample of 20 Year 8 students.

Solution

Find the number of Year 8 students.

$$\text{Number of Year 8s} = 395 - 247 + 1 = 149$$

Find the repeat for selection.

$$149 \div 20 \approx 7$$

Pick a place to start at random.

Start with student number 328.

Choose every 7th student. When you get to the end of the list, still count in 7s into the start of the list.

The students chosen should be those with library IDs 328, 335, 342, 349, 356, 363, 370, 377, 384, 391, 249, 256, 263, 270, 277, 284, 291, 298, 305 and 312.

While the best samples are generally considered to be random samples, they can be difficult and expensive to use. For a random sample to be valid, you need to make sure that the whole selected sample is used. When people are involved, this can be difficult, as some may not be available when you try to obtain responses. You cannot ignore people who are absent as you could end up excluding a particular group when you only do surveys at particular times, and this could cause significant bias. In practice, cost dictates the number of times you try to include responses from absent selections.

Random sampling of people is difficult and expensive, so surveys may be conducted using non-random samples. You might just choose a convenient group, such as people who live close by; you might try to include people you think would be representative of the population; you might decide to include people who attend a particular event; or you might know that 6% of the adult population are unemployed, 28% do not work (retired, disabled, etc.) and the rest are working so you choose quotas of 3 unemployed people, 14 non-workers and 33 workers from a crowd of people.

IMPORTANT

Convenience sampling is choosing from a convenient group of the population.

Judgment sampling is the use of judgement to determine a representative sample.

Purposive sampling is choosing representatives that meet particular conditions.

Quota sampling is choosing the first convenient representatives for a sample according to the proportions of particular divisions of the population, such as male and female.

If you were choosing a sample from the company staff in Example 5 and you were in the office, then just using office staff would be a convenient sample. If you decided on people you thought would make the best sample, it would be judgement sampling. If you decided to choose people who had been employed for more than 5 years at the company, it would be purposive sampling and if you just used the first 4 office staff, the first 4 drivers and the first 2 store staff you came across, it would be quota sampling.

INVESTIGATION**Random and non-random samples**

Use newspapers, magazines, TV reports, advertisements and internet reports of surveys.

Classify the surveys as random or non-random and then further classify them into types.

Order the surveys so that they show how the samples used are representative of the population involved. Write a report of your findings.

EXERCISE 9.02 Selection of samples

Concepts and techniques

- 1 **Example 4** Start at row 8, column 17 of the 2-digit random number table on page 468 and select 8 different numbers that are:
 - a 2-digit numbers between 20 and 99
 - b 6-digit numbers between 100 000 and 400 000
 - c 1-digit numbers
 - d 2-digit numbers between 30 and 84.
- 2 Start at row 9, column 25 of the 2-digit random number table on page 468 and select:

a 10 different numbers from 1 to 50	b 6 different numbers from 450 to 700
c 8 different numbers from 1500 to 2500	d 12 different 3-digit numbers.
- 3 **CAS** Select 8 random integers from 28 to 2198.
- 4 **Example 5** How many of each group should be selected to make stratified random samples from each of the following?
 - a A sample of 10 swimsuit-wearers from 15 men in board shorts, 10 men in briefs, 25 women in bikinis and 7 women in one-pieces.
 - b A sample of 20 chocolates from 180 soft-centred, 140 hard-centred, 85 liquid-centred and 108 nutty-centred chocolates.
 - c A sample of 16 from 5 fifteen-year-olds, 35 sixteen-year-olds, 10 seventeen-year-olds and 3 eighteen-year-olds.
 - d A sample of 15 staff from a manufacturing firm that employs 42 assembly workers, 10 office staff and 3 supervisors.
- 5 **Example 6** Starting from the given number, state which numbers to use for systematic samples for each of the following.
 - a A sample of 10, starting at 28 from a group numbered from 5 to 146.
 - b A sample of 8, starting at 216 from a group numbered from 105 to 327.
 - c A sample of 15, starting at 64 from a group numbered from 1 to 427.
 - d A sample of 9, starting at 1472 from a group numbered from 1257 to 2832.

- 6 State the kind of sampling used in each of the following non-random samples of students at a school.
- Peter asked the first 10 girls and 10 boys who came into the library
 - Vera asked 10 of her friends
 - David asked students he thought would be good representatives
 - Sally interviewed the students who played netball
 - Ami interviewed the students on her bus on the way to school
 - Michael asked students at lunchtime until he had 3 from each of Years 7 to 12
 - Celia thought people on the school council would be best, so chose 6 of them in Years 11 and 12
 - Corey chose people who had been at the school since Grade 7
- 7 Use systematic sampling to select 10 students using the enrolment numbers at a school. Start from enrolment number 5130 of the current enrolments, which run from 4928 to 5672 inclusive.

Reasoning and communication

- 8 Biotech Industries Pty Ltd wishes to form a staff social committee consisting of 18 members. The firm decides to use the method of stratified random sampling for selecting the committee members. The employment details of the firm are given in the table.

	Administrative staff	Factory workers
Males	11	73
Females	24	52

- How many from each group should be selected to represent all groups fairly?
 - How many from each group should be selected if no distinction is made between males and females?
- 9 The table below shows the Australian population in 2012. Use stratified random sampling to determine how many should be chosen from each state to make a sample of 500:
- if persons are selected regardless of sex
 - if males and females are selected in proportion.

Australian population, September 2012

State	Males	Females	Persons
New South Wales	3 628 553	3 685 546	7 314 099
Victoria	2 793 330	2 855 722	5 649 052
Queensland	2 285 309	2 299 280	4 584 589
South Australia	820 740	837 408	1 658 148
Western Australia	1 236 308	1 215 137	2 451 445
Tasmania	255 169	257 006	512 175
Northern Territory	124 080	112 269	236 349
Australian Capital Territory	187 329	189 131	376 460
Australia	11 332 884	11 452 580	22 785 464

Source: ABS (3101.0 Table 4)

- 10 O’Hea Street in Coburg has numbers from 1 to 386. Use systematic sampling to choose 20 houses for an employment survey, starting from 205 O’Hea St.

9.03 VARIABILITY OF RANDOM SAMPLES

Unless you use the whole population, random samples will vary. For a small population like {3, 5, 7, 8, 11} you could take samples of 2 items. Some of those samples would be {3, 5}, {3, 11} and {7, 8}. These are clearly very different and it is obvious that samples as small as this might give you a very misleading idea about the population distribution.



Sample generator

INVESTIGATION Sample variation

You can use the Excel spreadsheet *Sample generator* on the website to generate samples from uniform, normal, binomial or Bernoulli distributions to compare them. The spreadsheet also calculates the mean and standard deviation of the sample.

Sample generator

Instructions

Use the first spinner to choose the distribution: uniform, normal, binomial or Bernoulli.

Type in the parameters for the distribution. A maximum of 50 trials is available for a binomial distribution.

Use the second spinner to choose the number in the sample up to a maximum of 200.

Click on the 'Get Sample' button to get a sample. Click again to get another sample.

Type of distribution	Lower boundary	Upper boundary	Number in sample
<input type="button" value="▲"/> Uniform <input type="button" value="▼"/>	50	80	<input type="button" value="▲"/> 30 <input type="button" value="▼"/>

Get Sample

Sample mean	Sample standard deviation
64.9793065	8.638457654

Sample

53.8187
56.1604
62.4938
...

- Use the spreadsheet to generate a sample of 20 numbers from a uniform distribution on the interval 50–80.
- Draw a graph of the sample.
- Now get another sample and draw a graph of this sample.
- How do the graphs compare to each other?
- Compare box-and-whisker plots of the samples.
- Get another two samples and draw box-and-whisker plots of the samples.

- Now get 4 samples of 20 values from the binomial distribution with $n = 30$ trials and probability of success $p = 0.4$.
- Compare box-and-whisker plots of your binomial samples.
- Get 4 samples of 20 values from the normal distribution with mean $\mu = 40$ and standard deviation $\sigma = 8$.
- Compare box-and-whisker plots of your normal samples.

You will have seen from the investigation that each sample is likely to be different. Their means and standard deviations are also different, but the larger the samples, the more similar they are likely to be.

Example 7

CAS Two random samples were taken from a Bernoulli distribution with probability $p = 0.42$. Each sample has 20 values. The samples are shown below.

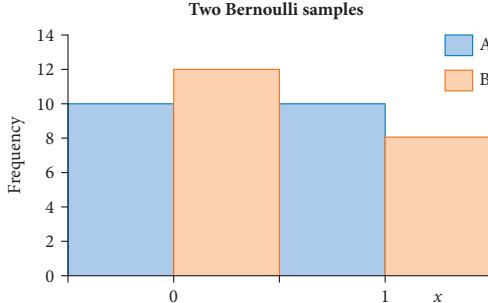
A: 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1

B: 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0

- Show the samples as a side-by-side column graph.
- Calculate the means and standard deviations.
- Compare the samples.

Solution

- a Draw the graphs with different colours for each sample.



- b Calculate the mean and standard deviation of each sample.

$$\mu_A = 0.5, \sigma_A = 0.5$$

$$\mu_B = 0.4, \sigma_B \approx 0.49$$

Compare the samples.

Sample A has a higher mean than sample B, and this is evident on the graph. However, their spreads are similar, as shown on the graph, and by their standard deviations.

Your CAS calculator may be used to generate a list of random numbers between 0 and 1 and you can manipulate this list to produce a random sample from a uniform distribution, for a random number $0 < x < 1$, $a < a + (b - a)x < b$.

Example 8

CAS a Generate two samples of 15 items from a uniform distribution on the interval $[30, 50]$, placing each sample in a different column of a spreadsheet.

b Draw a graph of the first distribution

c Compare to the second distribution and comment on the appearance.

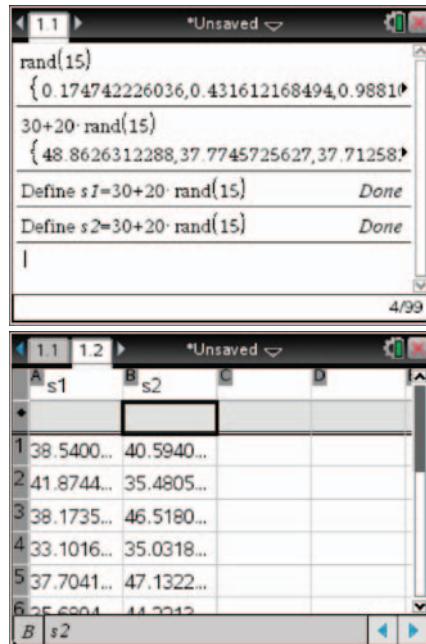
Solution

TI-Nspire CAS

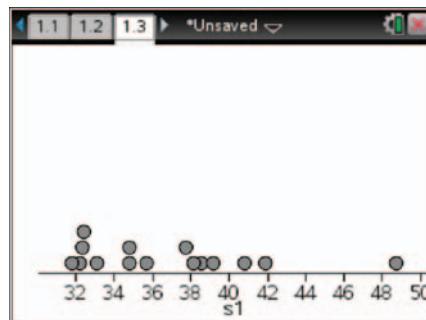
a $\text{rand}(15)$ will produce 15 random numbers, so $30 + 20 \times \text{rand}(15)$ produces a sample of 15 items from the required distribution.

Define two such samples as $s1$ and $s2$.

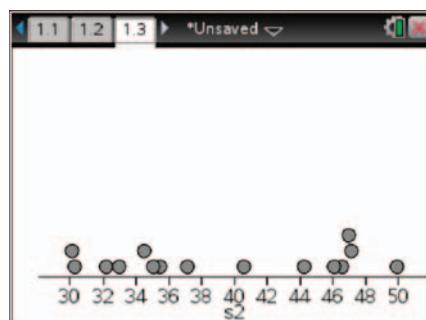
You can easily examine the lists by inserting a Lists & Spreadsheet page. Type the variables $s1$ and $s2$ into the column headings and they will be copied into the columns.



b Insert a Data & Statistics page and ‘Click to add variable’ at the bottom. Choose $s1$. A dot graph of the sample will be shown.



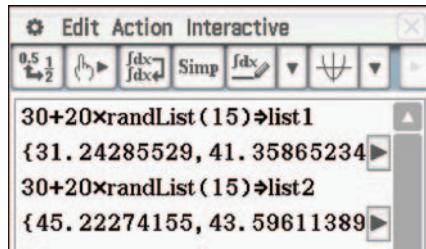
c Click on the variable $s1$ at the bottom of the graph and choose $s2$ instead. The display will change to a dot plot of $s2$.



ClassPad

- a Make sure the calculator is set to decimal. $\text{randList}(15)$ will produce 15 decimal numbers between 0 and 1, so $\text{randList}(15) \times 20 + 30$ will produce 15 decimal numbers between 30 and 50. Name this list1. Repeat for list2. A different set of random numbers will be produced.

You can see both distributions in the Statistics menu.



The calculator screen shows two lists generated by the ClassPad:

```
30+20×randList(15)⇒list1
{31.24285529, 41.35865234}
30+20×randList(15)⇒list2
{45.22274155, 43.59611389}
```

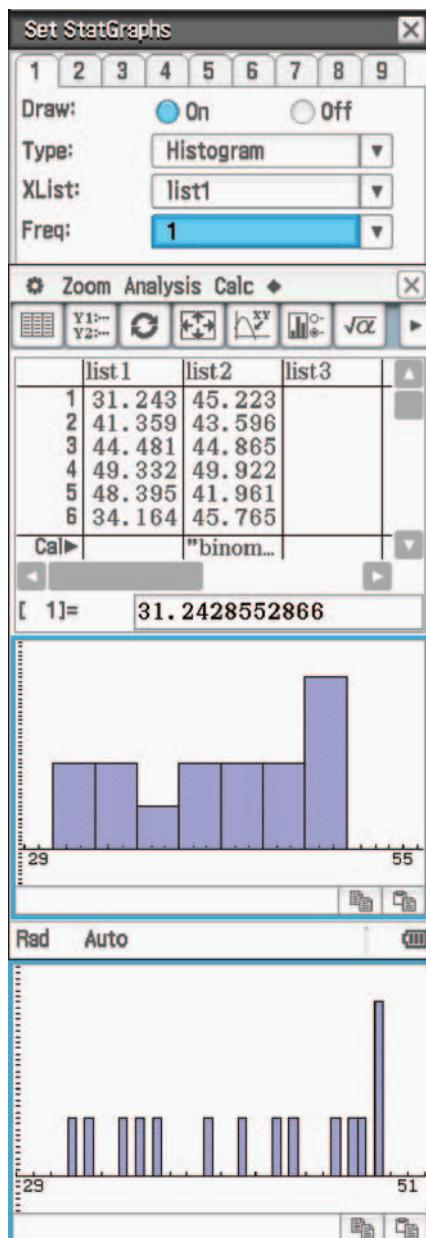
- b Go to the Statistics menu.

Tap SetGraph and Setting. Choose as shown on the right.

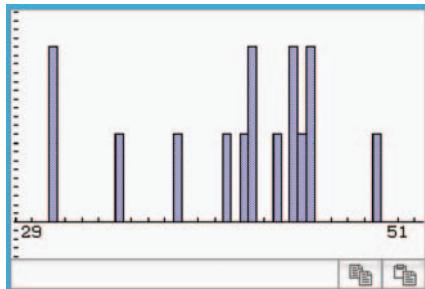
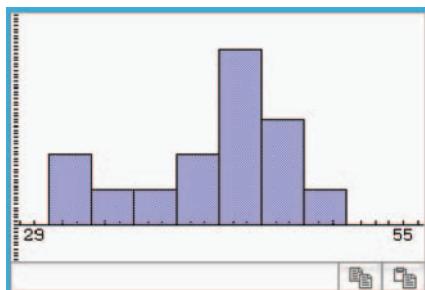
Tap to draw the graph. Make HStart 30 and HStep 3.

Tap the top half of the graph, and repeat using HStart 30 and HStep 0.5.

You will get a different type of graph.



- c Repeat for the second distribution. The only change will be that XList will now be list2 and not list1.
Only the graphs are shown on the right.



Compare the sample graphs.

The samples appear to be different.

In Example 8, each sample is different. If you follow the instructions for Example 8, your samples will almost certainly be different to those shown, but the general conclusions will be similar.

You can use `randNorm(μ , σ , #s)` on the TI-Nspire CAS or `randNorm(σ , μ , #s)` on the ClassPad to get a list of #s random values from a normal distribution with mean m and standard deviation s. Similarly, `randBin(n , p , #s)` gives a list of #s random values from a binomial distribution with n trials and probability p on both calculators. A Bernoulli distribution is a binomial distribution with one trial, so `randBin(1, p , #s)` gives a list of #s random values of 0 or 1 where success (1) has a probability p on both calculators. You can also find the mean and standard deviation of a list using 1-variable statistics.

Example 9

- CAS**
- Generate two random samples of 20 items from a normal distribution with mean $\mu = 25$ and standard deviation $\sigma = 6$.
 - Find the mean and standard deviation of each sample.
 - Compare the samples.

Solution

TI-Nspire CAS

- Define the random samples as s1 and s2 using the `randNorm` function.
You might want to look at the samples as columns in a spreadsheet of dot plots as shown in Example 7.

```
Define s1=randNorm{25,6,20}
Define s2=randNorm{25,6,20}
```

- b Press **[menu]**, 6: Statistics, 1: Stat Calculations and 1: One-Variable Statistics. Choose 2 lists and type s_1 and s_2 as the X1 List and X2 List.

Scroll the lists up and down and to the right to find the mean and standard deviations of the lists, shown as \bar{x} and $\sigma_x := \sigma_{nx}$.

Write the results.

OneVar 2,s1,s2: stat.results	
"Title"	"One-Variable Statistics"
" \bar{x} "	25.1925180145
" Σx "	503.850360291
" Σx^2 "	13420.1605962
" $s_x := \sigma_{nx}$ "	6.18530226461
" $\sigma_x := \sigma_{nx}$ "	6.02868691336
"n"	20
"MinX"	15.0453822891
"MaxX"	25.9546177109

$$\mu_{s1} \approx 25.19, \sigma_{s1} \approx 6.03, \mu_{s2} \approx 27.25, \sigma_{s2} \approx 7.09$$

- c Compare the samples.

The samples have different means, standard deviations and appearances.

ClassPad

- a Use the Catalog to select randNorm. Enter in the order standard deviation, mean, and number of random samples. Store the data in list1 and list2.

```
randNorm(6, 25, 20)⇒list1
{30.92564683, 28.11594012
randNorm(6, 25, 20)⇒list2
{25.49770268, 26.92371376
```

- b Go to the Statistics menu.

Tap Calc, One-Variable and set XList to list1 and Freq to 1.

Read the mean, $\bar{x} \approx 26.34$ and standard deviation $\sigma_x \approx 4.60$.

Repeat for list2.

Read the mean, $\bar{x} \approx 24.17$ and standard deviation $\sigma_x \approx 7.00$.

Stat Calculation	
One-Variable	
\bar{x}	=26.343962
Σx	=526.87923
Σx^2	=14303.424
σ_x	=4.6007474
s_x	=4.7202672
n	=20
minX	=17.377331
Q_1	=22.40694
Med	=26.720754
σ_x	=4.6007474

OK

Stat Calculation	
One-Variable	
\bar{x}	=24.170716
Σx	=483.41432
Σx^2	=12664.505
σ_x	=7.0001242
s_x	=7.1819759
n	=20
minX	=7.4043231
Q_1	=19.162855
Med	=25.00713
σ_x	=7.0001242

OK

$$\mu_{s1} \approx 26.34, \sigma_{s1} \approx 4.60, \mu_{s2} \approx 24.17, \sigma_{s2} \approx 7.00$$

- c Compare the samples.

The samples have different means, standard deviations and appearances.

EXERCISE 9.03 Variability of random samples

Concepts and techniques

- 1 **Example 7** Two random samples were taken from a Bernoulli distribution with $p = 0.6$. Each sample has 20 values. The samples are shown below.
A: 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1
B: 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1
a Represent the samples using a side-by-side column graph.
b Compare the samples.
- 2 Two random samples were taken from a uniform distribution on the interval [10, 25]. Each sample has 18 values. The samples, rounded correct to 1 decimal place, are shown below.
A: 14.1, 17.1, 19.5, 10.3, 16.8, 22.7, 23.8, 21.7, 22.2, 24.9, 14.2, 18.2, 16.4, 17.9, 17.8, 13.2, 21.7, 18.6
B: 12.6, 16.5, 12.2, 10.8, 21.6, 12.9, 21.3, 23.8, 11.9, 20.5, 21.8, 24.7, 23.1, 14.5, 20.7, 17.3, 13.7, 18.1
a Represent the samples using side-by-side histograms with class widths of 2.
b Compare the samples.
- 3 Two random samples were taken from a normal distribution with a mean of 50 and a standard deviation of 8. Each sample has 25 values. The samples, rounded correct to 1 decimal place, are shown below.
A: 50.4, 62.3, 49.0, 45.3, 59.1, 45.1, 50.2, 49.7, 40.0, 41.6, 38.5, 46.6, 47.5, 47.1, 63.2, 55.4, 43.0, 52.1, 53.3, 54.9, 58.4, 34.2, 54.4, 37.1, 56.9
B: 37.3, 49.3, 35.9, 45.7, 53.5, 40.2, 46.7, 44.8, 52.8, 41.7, 59.4, 48.7, 50.2, 39.4, 50.3, 41.6, 43.4, 41.9, 46.1, 47.8, 45.1, 53.3, 44.9, 44.5, 50.6
a Represent the samples using side-by-side histograms with class widths of 5.
b Compare the samples.
- 4 **Examples 8, 9** **CAS** a Generate two random samples of 25 items from a binomial distribution with $n = 30$ trials and probability $p = 0.7$ and draw dot plots of the samples.
b Find the mean and standard deviations of the samples.
c Compare the samples.
- 5 a Generate two random samples of 20 items from a Bernoulli distribution with probability $p = 0.25$.
b Find the mean and standard deviations of the samples.
c Compare the samples.
- 6 a Generate two random samples of 30 items from a uniform distribution on the interval [5, 25] and draw dot plots of the samples.
b Find the mean and standard deviations of the samples.
c Compare the samples.
- 7 a Generate two random samples of 35 items from a normal distribution with mean $\mu = 50$ and standard deviation $\sigma = 10$ and draw dot plots of the samples.
b Find the mean and standard deviations of the samples.
c Compare the samples.

Reasoning and communication

- 8 a Generate two random samples of 9 items from a binomial distribution with $n = 15$ trials and probability $p = 0.6$ and find the mean and standard deviations of the samples.
b Generate two random samples of 64 items from a binomial distribution with $n = 15$ trials and probability $p = 0.6$ and find the mean and standard deviations of the samples.
c Comment on the variation of the means and standard deviations for samples of 9 and samples of 64 items.
- 9 a Generate two random samples of 16 items from a uniform distribution on the interval $[35, 45]$ and find the mean and standard deviations of the samples.
b Generate two random samples of 64 items from a uniform distribution on the interval $[35, 45]$ and find the mean and standard deviations of the samples.
c Comment on the variation of the means and standard deviations for samples of 16 and samples of 64 items.
- 10 a Generate two random samples of 9 items from a normal distribution with mean $\mu = 40$ and standard deviation $\sigma = 8$ and find the mean and standard deviations of the samples.
b Generate two random samples of 81 items from a normal distribution with mean $\mu = 40$ and standard deviation $\sigma = 8$ and find the mean and standard deviations of the samples.
c Comment on the variation of the means and standard deviations for samples of 9 and samples of 81 items.

9.04 SAMPLE PROPORTIONS

If you were checking whether Vitamin C supplements decrease the chances of catching a cold, you might ask one sample of people to take no vitamin C, another to take one capsule per day and a third sample to take 2 capsules per day. For each sample, the variable would be a random Bernoulli variable with success (if you can call it that) being ‘catching a cold’. You would then look at the frequencies of people in each group who caught colds. The actual frequencies would not be as important as the ratio of the numbers who caught colds to the numbers in the samples. Each ratio is an example of a **sample proportion**.

IMPORTANT

A **sample proportion** is the ratio of the number of times a property (characteristic) occurs in a sample, divided by the number in the sample. The sample proportion is denoted by \hat{p} (read as p hat).

The occurrence of the property is normally called a **success**, so for x successes in a sample of n , the sample proportion is given by $\hat{p} = \frac{x}{n}$.

○ Example 10

From 20 people who took no vitamin C, 8 got colds during one winter. What is the sample proportion of colds?

Solution

Success is getting a cold. Use the formula.

$$\hat{p} = \frac{x}{n}$$

Substitute $x = 8$ and $n = 20$ and calculate the answer.

$$= \frac{8}{20} = 0.4$$

Write the solution.

The sample proportion for those who got colds is 0.4.

You would expect the probability of success in a population to be close to the sample proportion. From the work you did in the last section, you would expect the variability of the sample proportion to decrease as the sample size increased. In the limiting case, when the sample is the whole population, it must be equal to the probability of success.

IMPORTANT

The sample proportion \hat{p} for the occurrence of a property (characteristic) in a population is an **estimator** of the probability p of the occurrence in the population.

○ Example 11

Police at a roadside checkpoint stopped 55 cars to check their roadworthiness. 7 of the drivers were issued with notices to have faults fixed within a week and 2 cars had such severe problems that they were immediately stopped from driving any further.

- Use the information to estimate the probability that a randomly selected car has a fault.
- State any problems with treating this as a reliable estimate.



Alamy/David Hancock

Solution

- a 9 out of 55 cars had faults affecting roadworthiness.

$$\text{Sample proportion} = \frac{9}{55} \approx 0.16$$

State the answer.

The estimated probability of faults is about 0.16.

- b Experienced police would be likely to stop cars they thought would be likely to have faults.

Since the police are unlikely to choose obviously new cars, the estimate may be higher than the true probability.

EXERCISE 9.04 Sample proportions

Concepts and techniques



Sample proportions

- 1 **Example 10** From a sample of 200 Irish High School students, 9 had red hair. What is the sample proportion of red hair for Irish High School students?
- 2 A normal die was tested by throwing it 120 times. It landed with 6 uppermost a total of 18 times. What is the sample proportion of '6' for this die?
- 3 A stove manufacturer checked stoves leaving the factory for faults. From 125 checked in one day, 8 were found to have faults in the paintwork that would make them 'factory seconds'. What was the sample proportion for factory seconds?
- 4 **Example 11** From a sample of 40 Australian Year 12 students, 9 were found to have heights of 180 cm or greater. Only one of the 9 was female.
 - a Estimate the probability of an Australian Year 12 student being 180 cm tall or greater.
 - b 19 of the 40 students were male. Estimate the probability of a male Australian Year 12 student being 180+ cm.
 - c Estimate the probability of a female Australian Year 12 student being 180+ cm.
- 5 From a sample of 14 male South African Year 12 students, 3 had heights of 180 cm or more. Estimate the probability of a male South African Year 12 student being 180+ cm.
- 6 A commercial art gallery had an exhibition of paintings with prices ranging from \$180 to \$7900, with a total of 74 paintings on show. 5 of the paintings were under \$400. Estimate the probability of paintings being priced under \$400.

Reasoning and communication

- 7 Are there any problems with the estimate in question 6?
- 8 The ages of 70 people towing caravans or driving motor homes on the Bruce Highway in Queensland in July were checked and 24 were found to be over 60 years old.
 - a Estimate the probability of caravan or motor home drivers being over 60.
 - b Are there likely to be any problems with this estimate?

9.05 PARAMETERS OF SAMPLE PROPORTIONS

How are sample proportions related to the probability of the property occurring in the population? The sample proportion you get for, say, blue eyes from a sample of 20 Australian Year 12 students, will not be the same for each sample. However, you would expect the mean of the sample proportions to be representative of the probability.

The probability of an Australian Year 12 student having blue eyes is 0.32. Since there are a very large number of Australian Year 12 students, the probability for a sample of 20 students will be the same for each student chosen for the sample. Each student chosen constitutes a Bernoulli trial because there are only two outcomes and the probability of blue eyes is the same for each. This means that for a sample of 20, the probability of success is 0.32 for each trial so the number of successes is a binomial distribution with $p = 0.32$ and $n = 20$. The number of blue-eyed students in the sample is a random variable, say B . You know from your work in Chapter 5 that

$$E(B) = np = 0.32 \times 20 = 6.4 \text{ and } \text{Var}(B) = npq = 20 \times 0.32 \times 0.68 = 4.352$$

For a linear transformation $Y = aX + b$ of a random variable X , you also know that

$$E(Y) = a \times E(X) + b \text{ and } \text{Var}(Y) = a^2 \times \text{Var}(X).$$

Consider the random variable $H = \frac{1}{20} \times B$.

Then $E(H) = \frac{1}{20} \times E(B) = \frac{1}{20} \times 6.4 = 0.32$ and $\text{Var}(H) = \left(\frac{1}{20}\right)^2 \times \text{Var}(B) = 0.01088$

But $\hat{p} = \frac{x}{n} = \frac{b}{20}$, so H is the random variable for the sample proportion \hat{p} of blue eyes in a sample of 20 Australian Year 12 students.

This implies that the mean of the sample proportion is 0.32 and the variance is 0.01088.

IMPORTANT

For samples that are small compared to the population, the sample proportion is effectively a random binomial variable.

If the probability of a particular property is p , then for samples of n items

$$E(\hat{p}) = p, \text{Var}(\hat{p}) = \frac{pq}{n} \text{ and } \text{SD}(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}.$$

You can prove that the parameters of sample proportions have the values above by using the same method as shown for the example of blue eyes in Year 12 Australian students.

Since $E(\hat{p}) = p$, \hat{p} is the best estimator of p .

○ Example 12

The probability of a normal car tyre lasting more than 60 000 km is about 0.34. What is the variance and standard deviation of the proportion of samples of 30 such tyres lasting more than 60 000 km?

Solution

Write down the parameters for \hat{p} .

$$p = 0.34, q = 0.66, n = 30$$

Write the formula for $\text{Var}(\hat{p})$.

$$\text{Var}(\hat{p}) = \frac{pq}{n}$$

Substitute values and calculate the answer.

$$= \frac{0.34 \times 0.66}{30}$$
$$= 0.00748$$

Find $\text{SD}(\hat{p})$.

$$\text{SD}(\hat{p}) = \sqrt{\text{Var}(\hat{p})}$$
$$\approx 0.0865$$

Write the answer.

The variance is 0.007 48 and the standard deviation is about 0.0865.

○ Example 13

A coin is tossed 20 times and the number of heads is noted. This experiment is repeated many times. What is the expected value and standard deviation of the sample proportion of heads?

Solution

Write down the parameters for \hat{p} .

$$p = 0.5, q = 0.5, n = 20$$

Write the formulas for $E(\hat{p})$ and $\text{SD}(\hat{p})$.

$$E(\hat{p}) = p \text{ and } \text{SD}(\hat{p}) = \sqrt{\frac{pq}{n}}$$

Substitute values.

$$E(\hat{p}) = 0.5 \text{ and } \text{SD}(\hat{p}) = \sqrt{\frac{0.5 \times 0.5}{20}}$$

Simplify $\text{SD}(\hat{p})$.

$$\approx 0.11$$

Write the answer.

The mean proportion of heads would be 0.5 with a standard deviation of 0.11.

○ Example 14

A class of Year 7 students investigated the results of dealing a card from a well-shuffled pack and checking its suit. Each student in the class dealt a card 50 times, replacing and shuffling the cards before dealing the next one. They each counted the number of times the card was a heart and recorded the proportion of times out of 50 as a decimal. What would be the mean and standard deviation of these results?

Solution

Write down the parameters for \hat{p} .

$$p = 0.25, q = 0.75, n = 50$$

Write the formulas for $E(\hat{p})$ and $SD(\hat{p})$.

$$E(\hat{p}) = p \text{ and } SD(\hat{p}) = \sqrt{\frac{pq}{n}}$$

Substitute in the values.

$$E(\hat{p}) = 0.25 \text{ and } SD(\hat{p}) = \sqrt{\frac{0.25 \times 0.75}{50}}$$

Simplify $SD(\hat{p})$.

$$\approx 0.061$$

Write the answer.

The mean and standard deviation of the results would be about 0.25 and 0.061 respectively.

EXERCISE 9.05 Parameters of sample proportions



WS

Sample proportion calculations

Concepts and techniques

- 1 **Example 12** Find the mean, variance and standard deviations of sample proportions for samples with the following probabilities and number in each sample.
a $p = 0.2$ and $n = 50$ **b** $p = 0.7$ and $n = 25$
c $p = 0.81$ and $n = 120$ **d** $p = 0.22$ and $n = 80$
- 2 **Example 13** A normal die is tossed and the number it lands on is noted. Samples of 45 tosses are taken and in each case the proportion of times that the number is less than 3 is calculated. What is the mean and standard deviation of the sample proportion when this experiment is repeated multiple times?
- 3 45% of Canadian high school students catch a bus to get to school. Samples of 200 students from high schools across Canada are surveyed to determine the proportion travelling by bus to school. What is the expected proportion and standard deviation of the sample proportion?

Reasoning and communication

- 4 **Example 14** A normal pair of dice is thrown and the total is noted. Samples of 30 such throws are performed and the proportion of times the total is more than 9 is calculated for each sample. What is the mean and standard deviation of the sample proportions for totals more than 9?

- 5 A card is cut from a well shuffled deck. This is done 100 times and the number of times that a picture card appears is noted in each case. When this is repeated many times, what is the mean and standard deviation of the sample proportion of picture cards?
- 6 A weighted coin was tossed 25 times and the proportion of heads was noted. This was done a total of 40 times and the mean number of heads was found to be 15.02 and the standard deviation of the sample proportion of heads was 0.1. Estimate the probability of heads for this coin.
- 7 Show that for a property with a probability of p , the mean sample proportion of samples of size n is $E(\hat{p}) = p$.
- 8 Show that for a characteristic with a probability of p , the variance of sample proportions of samples of size n is $\text{Var}(\hat{p}) = \frac{pq}{n}$.

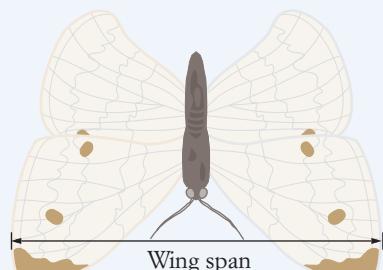
9.06 THE CENTRAL LIMIT THEOREM

You can calculate the mean and standard deviation of random variables from different samples. As you have seen already, these vary between samples, but as the sample size increases, the variation decreases and they become closer to the population mean and standard deviation. What happens with sample proportions?

INVESTIGATION Cabbage moths

Colin has a patch of cabbages, but the patch has attracted cabbage moths. He has used a large net to catch all the cabbage moths, so it is possible to work out the average size of the cabbage moths. Work in groups of about five people, so there are 5 groups in the class. Your teacher will give you some paper models of the cabbage moths to check and measure.

- 1 Work out the mean and standard deviation of the wing span of a sample of 6 cabbage moths.
- 2 How many of your sample are blue? What proportion are blue?
- 3 Now use samples of 12 cabbage moths.
- 4 Find the mean, standard deviation and proportion for samples of 18 and 25.
- 5 Compare your results with those of other groups.
 - What happens to the mean as the sample size is increased?
 - What happens to the standard deviation as the sample size is increased?
 - What happens to the proportion as the sample size is increased?
- 6 Work out the class averages for samples of 6, 12, 18 and 25. What happens to the class average as the sample size is increased?



To understand what happens to a statistic as sample size increases, it is useful to examine the *distribution* of the statistic for multiple samples. The distribution of a statistic for many samples is called a **sampling distribution**.

Collection of real data for multiple samples is very time-consuming, so you will use a simulation. Consider samples of Australian high school students and the property of having blue eyes. For Australian students, the probability of having blue eyes is 0.32. The occurrence of blue eyes is a Bernoulli random variable with a probability of success of 0.32. A sample of 20 students corresponds to 20 trials. The number of successes is a value x of the random binomial variable X and the sample proportion is $\frac{x}{20}$.

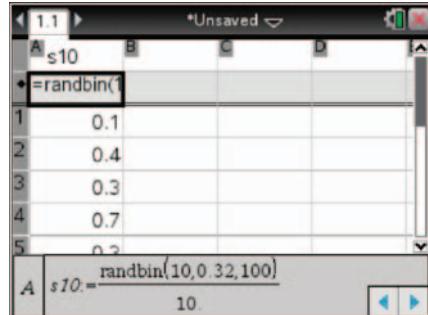
Example 15

- CAS**
- Simulate the sample proportion for blue eyes ($p = 0.32$) for 100 samples of 10 Australian students in the first column of a spreadsheet.
 - Create a dot plot of the distribution.
 - What shape does it appear to be?
 - Repeat the simulation for samples of 50 students in the next column and do a dot plot.
 - What happens to the shape of the distribution?

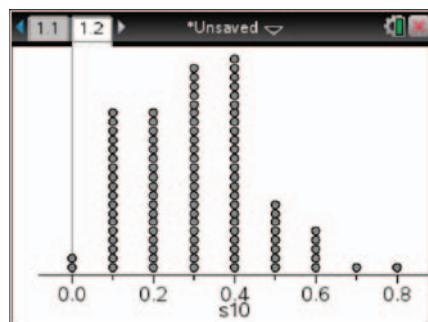
Solution

TI-Nspire CAS

- Use column A of a Lists & Spreadsheet page.
Type the variable name $s10$, say, into the top cell. Then type $\text{randBin}(10, 0.32, 100) \div 10$. into the next (formula) cell at the top. The simulated sample proportions for blue eyes for 100 samples of 20 students each will appear in cells A1–A100. The decimal point forces approximate calculation.



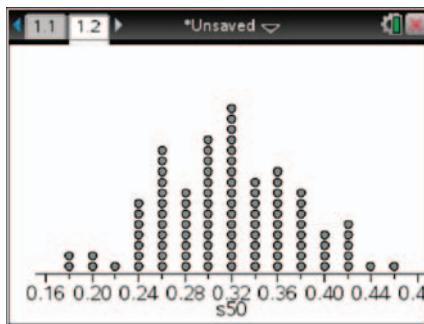
- Add a Data & Statistics page and click on the variable $s10$ at the bottom of the page.



- Comment on the shape.

The distribution is not symmetrical.

- d Repeat in column B but use the variable name $s50$ and the $\text{randBin}(50, 0.32, 30) \div 50$.
Click on $s50$ in the dot plot.

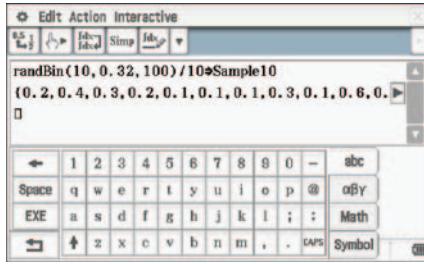


- e Comment on the change of distribution.

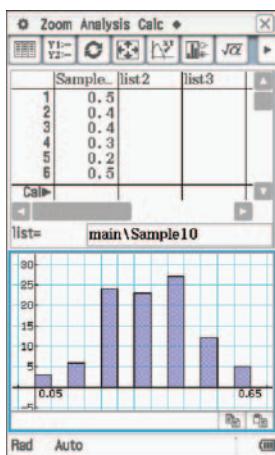
The distribution looks like a normal distribution.

ClassPad

- a Start at the Main menu. The randBin function generates a list of numbers and thus can be used to generate a list we shall call Sample10. Make sure the calculator is set to decimal.
It is not necessary to see all the numbers in the list, but you can do by tapping the arrow on the right of the numbers.



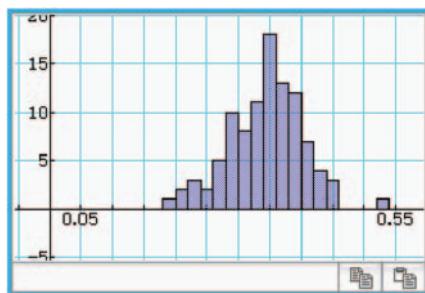
- b Go to the Statistics menu. Tap list1 and enter Sample10. If the numbers appear as fractions, tap the heading and then $\frac{\pi}{3.14}$. From SetGraph, tap Setting and choose Histogram, Main\Sample10 and 1. Use View Window to set the scale for x to 0.05 and the scale for y to 5. Tap the graph icon and set HStart to 0 and HStep to 0.05. For a graph without gaps, set HStep to 1.



- c Comment on the shape.

The distribution is not symmetrical.

- d Repeat but use the name Sample50 and $\text{randBin}(50, 0.32, 100) \div 100$. Use the second column and since the numbers change by a minimum of $\frac{1}{50} = 0.02$, make HStep 0.02.
The value falls on the left edge of each column.



- e Comment on the change of distribution.

The distribution looks more like a normal distribution with the mean a little less than 0.35. The graph roughly falls between 2 and 4.4, giving a mean of $\frac{2+4.4}{2} = 3.2$.

When the number in the sample is large, the distribution looks approximately normal. Remember that each time you get a sample, it is different, so your samples may appear a little different than those in Example 15, but the general pattern should be the same.

What happens to the sampling distribution for sample proportions as the value of p changes?

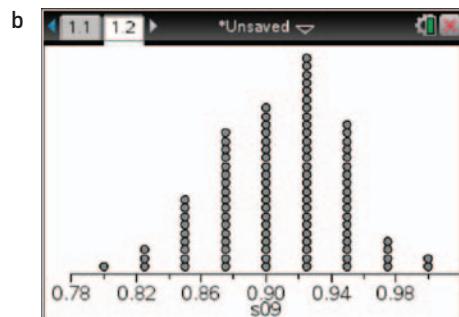
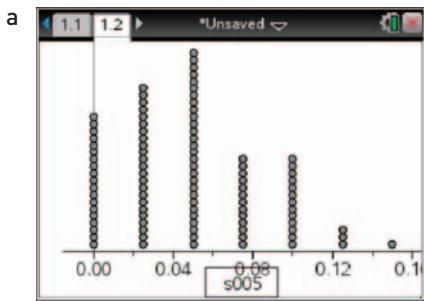
Example 16

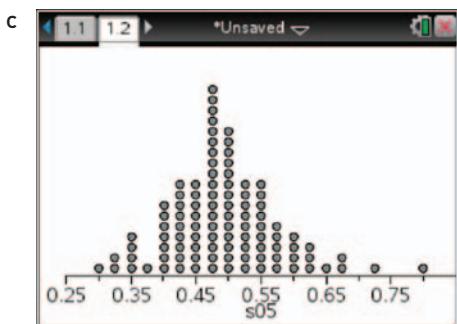
- CAS**
- a Simulate 100 samples of sample proportions with $p = 0.005$ and $n = 40$ and create a dot plot of the sampling distribution.
 - b Simulate 100 samples of sample proportions with $p = 0.9$ and $n = 40$ and do a dot plot of the sampling distribution.
 - c Simulate 100 samples of sample proportions with $p = 0.5$ and $n = 40$ and do a dot plot of the sampling distribution.
 - d Compare the sampling distributions.

Solution

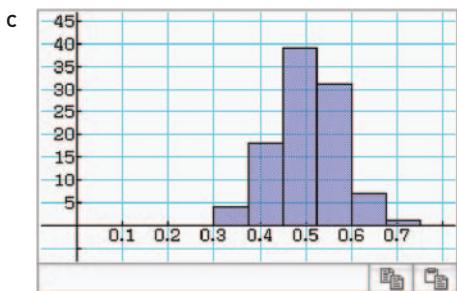
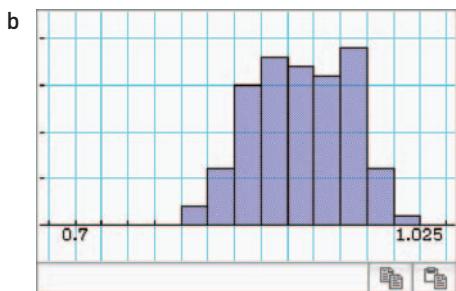
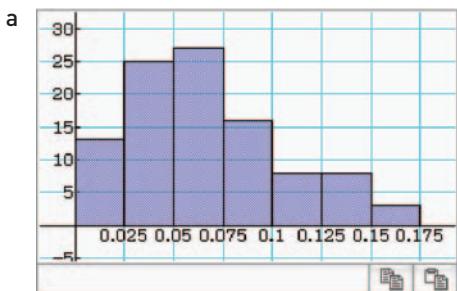
TI-Nspire CAS

Use $\text{randBin}(40, 0.005, 100) \div 40$ labelled as s005, $\text{randBin}(40, 0.9, 100) \div 40$ labelled as s09 and $\text{randBin}(40, 0.5, 100) \div 40$ labelled as s05.





ClassPad



- d Write a comment about the change in the shape of the distribution.

The distribution becomes more symmetrical and similar to a normal distribution as $p \rightarrow 0.5$.

When the probability of the property gets close to 0.5, the sampling distribution of the sample proportion approximates a normal distribution.

Both the number in a sample and the probability of the property for which the statistic is calculated affect the nature of the sampling distribution of a sample proportion. Statisticians use a guide to determine when the distribution of a sample proportion may be approximated by a normal distribution. This is important in applications because calculations involving the normal distribution are much simpler and quicker than those for a binomial distribution.

IMPORTANT

For $np > 5$ and $nq > 5$, the distribution of the sample proportion \hat{p} for a random Bernoulli variable is approximately normal with mean p and standard deviation $\sqrt{\frac{pq}{n}}$.

In fact, any binomial distribution for which $np > 5$ and $nq > 5$ may be approximated by a normal distribution with the same mean (np) and standard deviation (\sqrt{npq}).

What about other sampling distributions? The sample proportion is rather like the mean, which is the most common statistic in general distributions, along with the standard deviation.

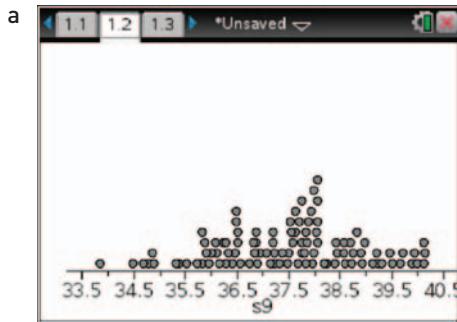
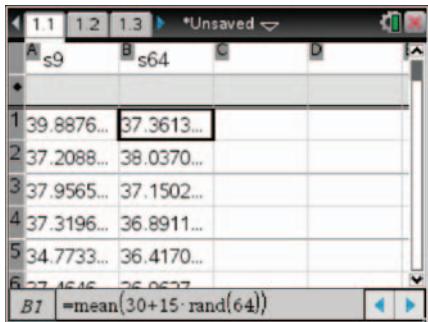
Example 17

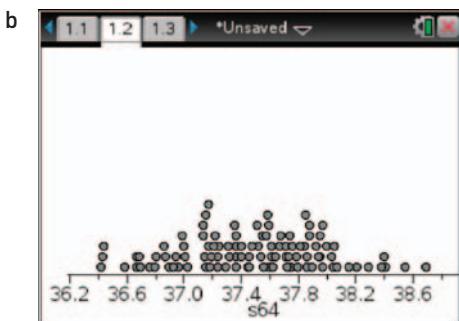
- CAS**
- a Simulate 100 samples of 9 items from a uniform distribution on the interval [30, 50] and draw a dot plot of the means of the samples.
 - b Simulate 100 samples of 64 items from a uniform distribution on the interval [30, 45] and draw a dot plot of the means of the samples.
 - c Compare the shapes of the sampling distributions.
 - d Compare the means and standard deviations of the sampling distributions.

Solution

TI-Nspire CAS

You can use $30 + 15 \times \text{rand}(9)$ to produce a sample of 9 items from the distribution (see Example 8). You can find the mean using $\text{mean}()$, so put the formulas $\text{mean}(10+15\times\text{rand}(9))$ and $\text{mean}(10+15\times\text{rand}(64))$ into cells A1 and B1 of a Lists & Spreadsheet page. Call the columns $s9$ and $s64$ and copy cells A1 and B1 down to cells A100 and B100. Add a Data & Statistics page and examine the dot plots.





ClassPad

Use the spreadsheet menu.

Tap Edit, Fill and Fill Range.

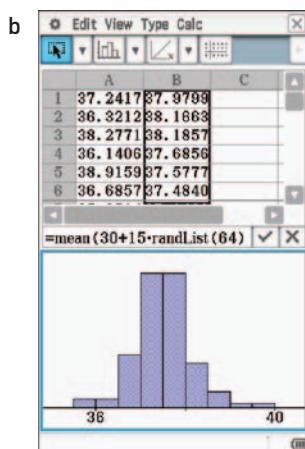
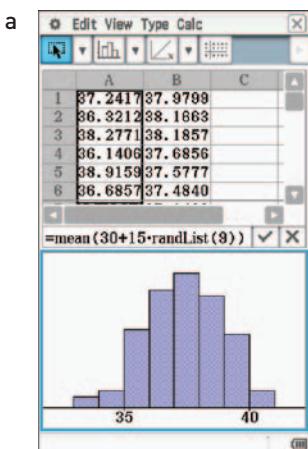
Enter the Formula $=\text{mean}(30+15\times\text{rand}(9))$ for the Range A1:A100. Note the symbols at the top of the calculator.

Repeat for column B, using the formula $=\text{mean}(30+15\times\text{randList}(64))$, and Range B1:B100.

Tap Edit, then Select, then Select Range, and enter A1:A100.

Tap Graph and then Histogram to draw a histogram. The ClassPad does not draw dot plots.

Repeat for a histogram of column B.



- c Comment on the shapes of the distributions.

The sampling distributions do not look like uniform distributions. They are more like normal distributions, with the larger sample even more like a normal distribution.

TI-Nspire CAS

- d Insert a Calculator page and find the 1-variable statistics for s_9 and s_{64} as in Example 9.

	1.1	1.2	1.3	*Unsaved	Variables	Equations	Calculus	Geometry	Statistics	Tables	Graphs	Notes	Links	Help
\bar{x}	37.4682921644	37.4												
Σx	3746.82921644	374												
Σx^2	140583.619172	140												
$= \bar{x} - x$	1.40822762938	0.49												
$= \sigma_x$	1.40116879982	0.49												
n	100													
$\ln x$	33.8394495582	36.4												
μx	36.4728668445	37.1												
$\text{mean } x$	37.5988233715	37.5												

Write the means and standard deviations.

$$\bar{x}_9 \approx 37.468, s_9 \approx 1.40, \bar{x}_{64} \approx 37.475, s_{64} \approx 0.490$$

ClassPad

Use the catalog to enter the following.

```
C1 =mean(A1:A100)  
C2 =stdDev(A1:A100)  
C3 =mean(B1:B100)  
C4 =stdDev(B1:B100)
```

A	B	C
39.448137	37.024137	55.570
37.730436	65.611	291.172
37.825337	58.9537	53.28
37.484437	42.31	0.59130
37.147236	36.7946	
36.962836	4125	

Write the means and standard deviations.

$$\bar{x}_9 \approx 37.557, s_9 \approx 1.29, \bar{x}_{64} \approx 37.533, s_{64} \approx 0.591$$

Write a comment.

The means are close to the mean of the original distribution (37.5), but the standard deviations are smaller than the original (about 4.3).

The uniform distribution and the normal distribution are very different from each other. However, the distribution of the sampling distribution of the means of a uniform distribution is similar to a normal distribution. This is true for all sampling distributions of means, no matter what the original distribution. As the number in a random sample increases, the similarity to a normal distribution increases. This is called the **central limit theorem**.

IMPORTANT

The **central limit theorem** states that for ‘relatively large’ random samples of a random variable X from a distribution with a finite mean μ and a finite standard deviation σ , the sampling distribution of the means is approximately normal. The approximation is also better for larger samples.

It can also be shown that $\bar{X} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, where n is the number in the sample.

A random sample means that the values of X are *independent*.

EXERCISE 9.06 The central limit theorem

Concepts and techniques

- 1 **Example 16** **CAS** a Simulate 100 samples of sample proportions with $p = 0.1$ and $n = 20$ and create a dot plot of the sampling distribution.
b Simulate 100 samples of sample proportions with $p = 0.1$ and $n = 50$ and create a dot plot of the sampling distribution.
c Simulate 100 samples of sample proportions with $p = 0.1$ and $n = 200$ and create a dot plot of the sampling distribution.
d Compare the sampling distributions.
- 2 **CAS** a Simulate 100 samples of sample proportions with $p = 0.8$ and $n = 20$ and create a dot plot of the sampling distribution.
b Simulate 100 samples of sample proportions with $p = 0.8$ and $n = 50$ and create a dot plot of the sampling distribution.
c Simulate 100 samples of sample proportions with $p = 0.8$ and $n = 200$ and create a dot plot of the sampling distribution.
d Compare the sampling distributions.
- 3 **CAS** a Simulate 100 samples of sample proportions with $p = 0.1$ and $n = 30$ and create a dot plot of the sampling distribution.
b Simulate 100 samples of sample proportions with $p = 0.3$ and $n = 30$ and create a dot plot of the sampling distribution.
c Simulate 100 samples of sample proportions with $p = 0.5$ and $n = 30$ and create a dot plot of the sampling distribution.
d Compare the sampling distributions.
- 4 **CAS** a Simulate 100 samples of sample proportions with $p = 0.08$ and $n = 20$ and create a dot plot of the sampling distribution.
b Simulate 100 samples of sample proportions with $p = 0.25$ and $n = 50$ and create a dot plot of the sampling distribution.
c Simulate 100 samples of sample proportions with $p = 0.5$ and $n = 90$ and create a dot plot of the sampling distribution.
d Compare the sampling distributions.
- 5 **Example 17** **CAS** a Simulate 100 samples of 12 items from a uniform distribution on the interval $[10, 40]$ and draw a dot plot of the means of the samples.
b Simulate 100 samples of 80 items from a uniform distribution on the interval $[10, 40]$ and draw a dot plot of the means of the samples.
c Simulate 100 samples of 200 items from a uniform distribution on the interval $[10, 40]$ and draw a dot plot of the means of the samples.
d Compare the shapes, means and standard deviations of the sampling distributions with the original distribution.

- 6 **CAS** a Simulate 100 samples from a binomial distribution with $p = 0.2$ and $n = 10$ and draw a dot plot of the means of the samples.
b Simulate 100 samples from a binomial distribution with $p = 0.2$ and $n = 30$ and draw a dot plot of the means of the samples.
c Simulate 100 samples from a binomial distribution with $p = 0.2$ and $n = 100$ and draw a dot plot of the means of the samples.
d Compare the shapes, means and standard deviations of the sampling distributions and the original distribution.
- 7 **CAS** a Simulate 100 samples of 6 items from a normal distribution with $\mu = 55$ and $\sigma = 12$ and draw a dot plot of the means of the samples.
b Simulate 100 samples of 25 items from a normal distribution with $\mu = 55$ and $\sigma = 12$ and draw a dot plot of the means of the samples.
c Simulate 100 samples of 120 items from a normal distribution with $\mu = 55$ and $\sigma = 12$ and draw a dot plot of the means of the samples.
d Compare the shapes, means and standard deviations of the sampling distributions with the original distribution.

Reasoning and communication

- 8 **Example 15 CAS** The Bureau of Meteorology issues long-range forecasts of rain based on ocean temperatures. One June, they say that there is a 65% chance of above median rainfall in Victoria for the 3-month period July–September.
a Simulate the sample proportion for above median rainfall for 100 samples of 12 Victorian locations in the first column of a spreadsheet.
b Draw a dot plot of the distribution.
c What shape does it appear to be?
d Repeat the simulation for 100 samples of 80 Victorian locations in the next column and draw a dot plot.
e What happens to the shape of the distribution?
- 9 **CAS** In Australia, about 24% of high school students can speak more than one language.
a Simulate the sample proportion of high school students who can speak more than one language for 100 samples of 12 students in the first column of a spreadsheet.
b Draw a dot plot of the distribution.
c What shape does it appear to be?
d Repeat the simulation for 100 samples of 80 students in the next column and draw a dot plot.
e What happens to the shape of the distribution?
- 10 **CAS** The average height of Year 12 boys is reported to be 184 cm with a standard deviation of 8.9 cm.
a Simulate 100 random samples for 15 heights of Year 12 Australian boys and find the means of each sample.
b Draw a dot plot of the sample distribution of the means.
c What shape does it appear to be?
d Repeat the simulation for 100 samples of 50 in the next column and draw a dot plot.
e What happens to the shape and standard deviation of the distribution?

9.07 SAMPLE PROPORTIONS AND THE STANDARD NORMAL DISTRIBUTION

You have seen that the distribution of sample proportions gets closer to the normal distribution as $n \rightarrow \infty$ and $p \rightarrow 0.5$. What happens to the distribution of $\frac{\hat{p} - p}{\sigma_{\hat{p}}}$?

INVESTIGATION Normalised sample proportions

- Simulate the sample proportion of 100 samples of 12 items with $p = 0.6$ in the first column of a spreadsheet.
- Find the mean and standard deviation of the sampling distribution.
- In the second column of the spreadsheet, subtract the mean and divide by the standard deviation. This gives the normalised sample proportions (like z -scores).
- Draw a graph of the new distribution.
- Draw the standard normal distribution on the same graph.
- Repeat the whole process for 100 samples of 40 items with $p = 0.6$.
- Do it again for 100 samples with of 200 items and $p = 0.6$.
- What happens to the normalised graph as n is increased?

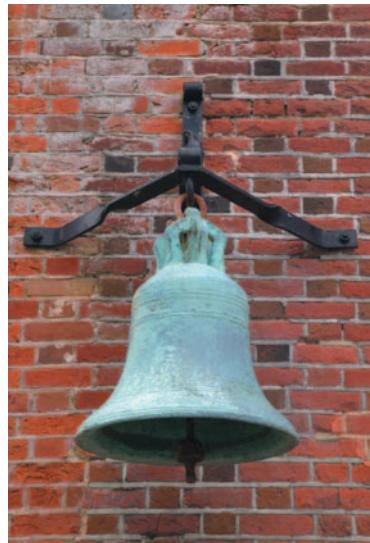
Whether or not a particular property occurs in a sample from a population is a Bernoulli variable, say X , which has the possible values 0 or 1. For x occurrences with a probability p from samples of n , the sample proportion $\hat{p} = \frac{x}{n}$ is the mean value \bar{X} .

The mean value of a binomial variable is np and the standard deviation is \sqrt{npq} . In this case, $n = 1$, so $\mu = p$ and $\sigma = \sqrt{pq} = \sqrt{p(1-p)}$.

According to the central limit theorem, the distribution of \bar{X} is approximately normal with $\bar{X} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, so $E(\hat{p}) = E(\bar{X}) = \mu = p$ and

$$SD(\hat{p}) = SD(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{pq}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{p(1-p)/n}.$$

This is exactly the same result as you saw in Section 9.05 for the parameters of \hat{p} .



Alamy / Rimage

In Chapter 8, you saw that any normal distribution can be transformed into a standard normal distribution using the linear transformation $Z = aX + b$, where $a = \frac{1}{\sigma}$ and $b = -\frac{\mu}{\sigma}$, so $Z = \frac{X - \mu}{\sigma}$.

For example, IQ has a mean of 100 and a standard deviation of 15. For an IQ of 130,

$Z = (130 - 100)/15 = 2$. It is 2 standard deviations above the mean. Applying this transformation to \hat{p} , the distribution approximates a standard normal distribution with $Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$.

IMPORTANT

For a Bernoulli variable X with parameter p , as $n \rightarrow \infty$, the distribution of $\frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$ approaches the standard normal distribution.

From the work you did in the last section, it should be clear that the closeness of the approximation to the standard normal distribution depends on the values of n and p . The larger the value of n and the closer p is to 0.5, the better the approximation.

Example 18

CAS What percentage of values of \hat{p} would lie between 0.45 and 0.55 for samples with $n = 80$ and $p = 0.4$?

Solution

Check the values of np and nq .

$$np = 80 \times 0.4 = 32$$

$$nq = 80 \times 0.6 = 48$$

Make a conclusion.

$np > 5$ and $nq > 5$, so the normal distribution can be used.

Find the mean and standard deviation.

$$\mu = p = 0.4$$

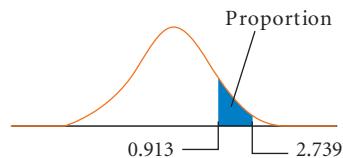
$$\text{SD}(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.4 \times 0.6}{80}} \approx 0.055$$

Find the standard values of \hat{p} .

$$\text{Standard value of } 0.45 = \frac{0.45 - 0.4}{0.055} \approx 0.913$$

$$\text{Standard value of } 0.55 = \frac{0.55 - 0.4}{0.055} \approx 2.739$$

Find the proportion of the standard normal distribution.



TI-Nspire CAS

The screenshot shows the TI-Nspire CAS software interface. In the input field, the command `normCdf(0.913, 2.739, 0, 1)` is entered, and the output is `0.177539869106`. The status bar at the bottom indicates `1/99`.

ClassPad

The screenshot shows the ClassPad software interface. It displays a series of calculations:
1. $\sqrt{\frac{0.4 \times 0.6}{80}}$ resulting in 0.05477225575
2. $(0.45 - 0.4)/s \Rightarrow p0$ resulting in 0.9128709292
3. $(0.55 - 0.4)/s \Rightarrow p1$ resulting in 2.738812788
4. `normCDF(p0, p1)` resulting in `0.1775702646`

Note that if a mean and standard deviation aren't given, `normCDF` treats it as a standard normal distribution.

Write the answer.

About 17.8% of the values would be between 0.45 and 0.55.

You can apply sample proportions to many situations involving samples.

Example 19

CAS Given that about 15% of Australians are left-handed, what is the probability that in a sample of 200 Australians, from 20 to 30 of them are left-handed?

Solution

Check np and nq .

$$np = 200 \times 0.15 = 30$$

$$nq = 200 \times 0.85 = 170$$

Write the conclusion.

$np > 5$ and $nq > 5$, so the normal distribution can be used.

Find the mean and standard deviation.

$$\mu = p = 0.15$$

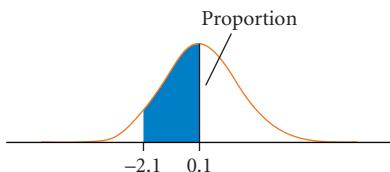
$$SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.15 \times 0.85}{200}} \approx 0.0252$$

Find the standard values of \hat{p} , using 19.5 to 30.5 to find the probability of the integers from 20 to 30 inclusive.

$$\text{Standard value of } \frac{19.5}{200} = \frac{0.0975 - 0.15}{0.025} \approx -2.1$$

$$\text{Standard value of } \frac{30.5}{200} = \frac{0.1525 - 0.15}{0.025} \approx 0.1$$

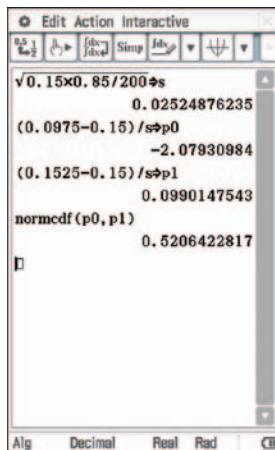
Find the proportion of the standard normal distribution.



TI-Nspire CAS



ClassPad



Write the answer.

The probability that 20 to 30 Australians from a sample of 200 are left-handed is about 0.52.

EXERCISE 9.07

Sample proportions and the standard normal distribution



Sample proportion probabilities

Concepts and techniques

- 1 **Example 18 CAS** What percentage of values of \hat{p} would lie between 0.5 and 0.6 for samples with $n = 30$ and $p = 0.7$?
- 2 **CAS** What percentage of values of \hat{p} would lie between 0.15 and 0.2 for samples with $n = 50$ and $p = 0.2$?
- 3 **CAS** What percentage of values of \hat{p} would lie between 0.75 and 0.85 for samples with $n = 100$ and $p = 0.9$?
- 4 **CAS** What proportion of values of \hat{p} would lie between 0.6 and 0.7 for samples with $n = 300$ and $p = 0.72$?
- 5 **CAS** What proportion of values of \hat{p} would lie between 0.48 and 0.52 for samples with $n = 45$ and $p = 0.25$?

Reasoning and communication

- 6 **Example 19 CAS** The probability of an adult having fair hair is 0.21. What is the probability that a sample of 400 adults has from 70 to 75 people with fair hair?
- 7 **CAS** About 25% of Australians are Victorians and of these, 76% live in Melbourne.
 - a From 300 Australians, what is the probability that from 50 to 60 are Victorian?
 - b From 300 Victorians, what is the probability that from 200 to 250 live in Melbourne?
 - c What is the probability that of the 300 Victorians, from 200 to 210 live in Melbourne?
- 8 **CAS** About 6.5% of Melbourne students travel to school by tram. What is the probability that from 50 to 60 Melbourne students from a random sample of 1000 travel to school by tram?

9

CHAPTER SUMMARY RANDOM SAMPLES AND PROPORTIONS

- For any variable or group of variables, the **population** is the whole group from which data could be collected.
- A **sample** is a part of the population.
- Data that is collected from the whole population is called a **census**.
- Data that is collected using a sample is called a **survey**.
- A **parameter** is a characteristic value of a particular population, such as the mean.
- A **statistic** is an estimate of a parameter obtained using a sample.
- A **fair sample** is one that is representative of the population.
- A **biased sample** is not representative: it favours some section of the population.
- A **random sample** is one chosen by a method that ensures that every member of the population has an equal chance of being chosen.
- **Selection bias** arises from the way the sample is chosen.
- **Design flaw bias** arises from faults in the design of a survey or census.
- **Interviewer bias** arises from differences in the way that interviewers seek information.
- **Recall/reporting bias** occurs when knowledge of the outcome of one answer affects recall or reporting of the answer of another.
- **Completion bias** occurs when surveys are incomplete.
- **Non-response bias** occurs when some subjects do not respond to the survey. **Self-selection bias** is an extreme form where subjects choose whether or not to take part.
- A **simple random sample** is one in which every member of the population has an equal chance of being selected.
- **Random numbers** are often used to choose a random sample after numbering a population.
- **Pseudo-random** numbers are generated by a rule that gives numbers that seem random.
- **Stratified random sampling** is a random choice of representatives from identifiable groups of the population, usually in proportion to the size of each group.
- **Cluster sampling** is a random choice of group(s) from the population as the sample.
- **Systematic sampling** is a choice of representatives from one group of the population by taking every n th one to make the desired sample size.
- **Convenience sampling** is a choice from a convenient group of the population.
- **Judgment sampling** involves the use of judgement to decide on a representative sample.
- **Purposive sampling** is a choice of representatives that meet particular conditions.
- **Quota sampling** is a choice of the first convenient representatives for a sample according to the proportions of particular divisions of the population, such as male and female.

- A **sample proportion** is the ratio $\hat{p} = \frac{x}{n}$ of the number of times x a property (characteristic) occurs in a sample, divided by the number n in the sample. The occurrence of the property is normally called a **success**. For samples that are small compared to the population, sample proportion is a random binomial variable.
- The sample proportion \hat{p} for the occurrence of a property in a population is an **estimator** of the probability p of the occurrence in the population.
- If the probability of a particular property is p , then for samples of n items

$$E(\hat{p}) = p, \text{Var}(\hat{p}) = \frac{pq}{n} \text{ and} \\ \text{SD}(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}.$$

- The distribution of a statistic for many samples is called a **sampling distribution**.

- For $np > 5$ and $nq > 5$, the distribution of the sample proportion \hat{p} for a random Bernoulli variable is approximately normal with mean p and standard deviation $\sqrt{\frac{pq}{n}}$.
- The **central limit theorem** states that for ‘relatively large’ random samples of a random variable X from a distribution with a finite mean m and a finite standard deviation s , the sampling distribution of the means is approximately normal. The approximation is also better for larger samples.
- It can also be shown that $\bar{X} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, where n is the number in the sample.
- A random sample means that the values of X are *independent*.
- For a Bernoulli variable X with parameter p , as $n \rightarrow \infty$, the distribution of $\frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$ approaches the standard normal distribution.

CHAPTER REVIEW

RANDOM SAMPLES AND PROPORTIONS

9

Multiple choice

CHAPTER REVIEW • 9

Short answer

- 7 **Example 1** Identify the population, some parameters and some statistics for each of the following.
- 40 people on a tram were asked which football team they most disliked. 10 said Collingwood, 4 said Hawthorn, 3 said Brisbane and the rest of the teams were chosen by at most 2 people.
 - From the odd-numbered houses on one side of a street, 2 have one bedroom, 12 have 2 bedrooms, 46 have 3 bedrooms, 9 have 4 bedrooms and 3 have 5 bedrooms.
- 8 **Example 3** Which statistic is easier to collect?
- The proportion of serious accidents that occur in DIY home renovation.
 - The proportion of DIY home renovators who have serious accidents.
- 9 **Example 4 CAS** Select 7 random integers between 38 and 240 inclusive.
- 10 **Example 5** A workshop has 28 welders, 14 boilermakers, 16 sheet metal workers and 40 labourers. How many of each should be selected for a stratified sample of 10 workers?
- 11 **Example 6** The account numbers for the customers of a business run from 10 000 to 21 314. Use systematic sampling to obtain a sample of 8 customer account numbers, starting from customer number 17 113.
- 12 **Example 7** Two random samples of 20 items each were taken from a binomial distribution with $n = 24$ and $p = 0.45$. The samples were as follows.
A: 14, 12, 16, 14, 13, 12, 14, 8, 17, 12, 11, 9, 12, 11, 8, 11, 13, 14, 13, 14
B: 14, 8, 12, 7, 9, 12, 5, 9, 9, 13, 12, 9, 12, 9, 10, 9, 11, 13, 9, 10
 - Draw a side-by-side column graph of the samples.
 - Compare the samples.
- 13 **Example 10** From 24 students in a Maths Methods class, 18 achieved a pass, C or better. What is the sample proportion of students at least passing?
- 14 **Example 11** A fashion shop in a large shopping centre had jeans on clearance at a sale. Of the 8 brands on sale, 3 brands were \$75–\$124, 4 were \$125–\$165 and one brand was \$210.
 - Estimate the probability of jeans at the shopping centre being priced under \$125.
 - Are there any problems with this estimate?
- 15 **Examples 12–14** The probability of a weighted coin landing with ‘heads’ up is 0.59. The coin is tossed 30 times. What is the expected proportion of heads and the standard deviation of this proportion?
- 16 **Examples 15–17**
 - What happens to the distribution of sample proportions for a particular value of p as n gets larger?
 - What happens to the distribution of sample proportions for a particular value of n as p gets further from 0.5?
- 17 **Examples 18,19 CAS** What percentage of sample proportions would be expected to lie between 0.4 and 0.6 for samples with $n = 40$ and $p = 0.55$?

Application

- 18 In each of the following cases, state the kind of sampling employed and whether or not it is fair. If it is biased, state the kind(s) of bias.
- A Year 11 student asks each of the people in his class what kind of mobile they have and how many SMSs they send each week to determine mobile phone use among students.
 - A market research company rings 100 phone numbers taken at random from the residential phone directory to ask whether they vacationed in Victoria, interstate or overseas last year as part of a study for the tourism industry in Victoria. They were also asked to give the reasons for their choices.
- 19 **CAS**
- Generate two samples of 20 items from a uniform distribution on the interval [15, 25].
 - Find the mean and standard deviations of each sample.
 - Compare the sample results with each other and the theoretical mean and standard deviation.
- 20 A case of oranges with cardboard ‘dimple spacers’ between each layer contains 63 oranges. The probability that an orange packed in this way is rotten is 0.04. A fruit shop bought 20 boxes of oranges. Assuming that the probability of a distribution of 20 boxes having any values outside 3σ from the mean is negligible, what is the maximum number of rotten oranges the fruit shop can expect to find in a box?
- 21 46% of Australian students travel to school by car, but only about 23% of students in Britain travel to school by car. What is the probability that between 20 and 30 students from 50 students travel by car in each country?



Practice quiz